GraPHFormer: A Multimodal Graph Persistent Homology Transformer for the Analysis of Neuroscience Morphologies

Supplementary Material

Abstract

Quantitative analysis of neural morphology is central to understanding circuit development, computation, and pathology. Current methods often analyze topology or graph structure in isolation. We introduce GraPHFormer, a multimodal architecture that combines topological and graph representations through contrastive learning. The vision branch processes a three-channel persistence image encoding unweighted, persistence-weighted, and radius-weighted densities via DINOv2-ViT-S. In parallel, a TreeLSTM encoder captures geometric and radial attributes from skeleton graphs. Both project to a shared embedding space trained with symmetric InfoNCE loss. We evaluate GraPHFormer on six benchmarks (BIL-6, ACT-4, JML-4 , N7, M1-Cell, M1-REG) under self-supervised and supervised settings, demonstrating consistent improvements over topology-only, graph-only, and morphometrics baselines. We further demonstrate practical utility through cross-domain transfer between neuronal and glial morphologies and embedding space analysis.

1. Overview

This supplementary document provides additional experimental details, ablation studies, and analyses that complement the main paper. Specifically, we present:

- **Additional ablation studies** (Section 2): We evaluate image encoder architectures, analyze the statistical independence of RGB channels through correlation analysis, and visualize individual channel contributions.
- **Cross-domain generalizability** (Section 3): We assess transfer learning between neuronal and glial morphologies across species, demonstrating that the learned representations capture organizational principles that generalize across cell classes.
- **Embedding space visualization** (Section 4): We present t-SNE projections of learned embeddings across multiple datasets, revealing clustering patterns that reflect morphological organization.
- **Cross-modal retrieval analysis** (Section 5): We evaluate bidirectional retrieval between tree graphs and persistence images, characterizing the alignment and information asymmetry between modalities.
- **Alternative training strategies** (Section 6): We explore MoCo-style training with various fusion mechanisms, comparing attention-based and simple fusion strategies.
- **Visual correspondence** (Section 7): We provide qualitative examples illustrating the relationship between tree structures and their persistence image representations.

These analyses provide deeper insights into architectural choices, learned representations, and cross-domain applicability.

2. Additional Ablation Studies

2.1. Image Encoder Architecture

Table 1 compares six image encoder architectures for neuron morphology representation learning, using TreeLSTM as the tree encoder across all experiments. We evaluate DinoV2-S, ResNet18, ResNet50, SmallViT, and two hybrid variants (R18-ViT and R50-ViT) that replace the last two ResNet layers with vision transformer blocks. Models are trained for 50 epochs with self-supervised learning and evaluated every 5 epochs using frozen k-NN classification. Results are averaged over 5 random seeds. DinoV2-S achieves the best average performance (70.6%), excelling on BIL (87.0%) and JML-4 (74.3%), while ResNet18 performs best on ACT (54.2%).

Table 1. Ablation study for image encoder selection across different architectures. Models are trained for 50 epochs using self-supervised learning and evaluated every 5 epochs with frozen k-NN classification on three benchmark datasets. Hybrid variants (R18-ViT, R50-ViT) replace the last two ResNet layers with ViT blocks. Results are averaged over 5 random seeds with best performance in **bold**.

Model	ACT-4	BIL-6	JML-4	Average
R18-ViT	49.7 ± 1.4	85.7 ± 0.0	72.6 ± 4.0	69.3 ± 1.8
R50-ViT	51.0 ± 1.4	84.9 ± 0.7	71.5 ± 0.7	69.1 ± 0.9
ResNet50	52.9 ± 3.3	84.5 ± 1.8	71.5 ± 2.1	69.7 ± 2.4
SmallViT	49.6 ± 1.5	86.0 ± 1.1	70.3 ± 2.5	68.6 ± 1.7
ResNet18	**54.2 ± 1.0**	84.9 ± 1.2	71.7 ± 1.7	70.3 ± 1.3
DinoV2-S	50.6 ± 0.9	**87.0 ± 2.1**	**74.3 ± 1.9**	**70.6 ± 1.6**

2.2. RGB Channel Independence Analysis

We validate our multi-channel encoding by extracting features from single-channel persistence images and computing pairwise Spearman correlations on BIL-6. Table 2 shows moderate correlations (0.35–0.45, average 0.41). The G–B pair (persistence versus radius) exhibits lowest correlation (0.354), demonstrating that branch length and

thickness encode complementary geometry. While not perfectly orthogonal ($\rho < 0.2$), moderate correlations are expected since all channels derive from the same underlying morphology. The ablation results presented in the main paper show that RGB (63.3%) substantially outperforms single ($\leq$60.7%) or dual ($\leq$61.7%) channel configurations, confirming that all three channels contribute unique information despite moderate correlation.

Table 2. Pairwise correlation between RGB channel features (BIL-6 dataset). Moderate correlations indicate related but distinct morphological aspects. The G–B pair (0.354) is most orthogonal, showing that persistence and radius capture complementary information.

Channel Pair	Correlation	Interpretation
R – G	0.433	Moderate
R – B	0.447	Moderate
G – B	0.354	Lower (more orthogonal)
Average	**0.411**	**Moderate independence**

Figure 1 visualizes the individual and combined RGB channel encodings. Each channel encodes distinct topological features: the R channel shows high intensity in regions corresponding to dense branching near the soma, while the G channel emphasizes persistent branches in intermediate regions. Channel combinations (RG, RB, GB) demonstrate how different pairings capture complementary information, with the full RGB representation providing comprehensive encoding of topological features.

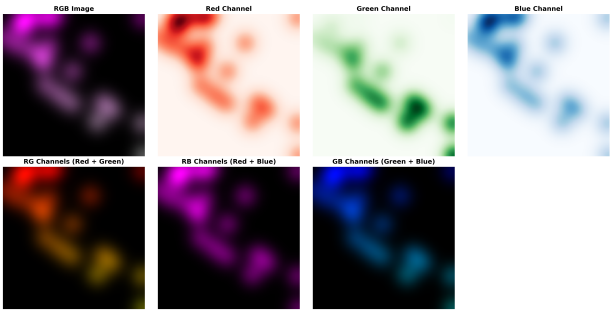

Figure 1. Visualization of RGB persistence image encoding and channel ablations. Each channel encodes distinct topological features. The R channel exhibits high intensity in the upper region while the G channel shows high intensity in the middle area. Channel combinations (RG, RB, GB) demonstrate how different pairings capture complementary information, with the full RGB representation providing comprehensive encoding of topological features.

2.3. Image resolution and Gaussian kernels.

We evaluated the impact of image resolution and Gaussian kernel size σ on model performance using supervised learning on the Neuron7 and JML datasets. Image resolutions ranged from 28 to 336 pixels (in multiples of 14 to accommodate the DinoV2 encoder), while kernel values σ ranged from 4 to 20. Results are presented in Figure 2. Performance peaked at resolution 112 for both datasets, achieving 71.53% on Neuron7 and 57.52% on JML-4, then declined at higher resolutions. For Gaussian kernels, optimal performance occurred at $\sigma = 16$ with 70.1% on Neuron7. Based on these results, we use an image resolution of $112 \times 112 \times 3$ and $\sigma = 16$ for all subsequent experiments.

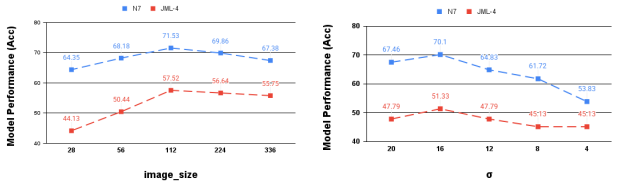

Figure 2. Ablation study on Image size (left) and Gaussian Kernel (right) using Image Encoder Only.

2.4. Embedding, Batch size and Projection Head

Embedding Size. We ablate the projection embedding dimension across $\{32, 64, 128, 256\}$, trained jointly on BIL-6, ACT-4, and JML-4 for 50 epochs under the self-supervised scheme. EM-128 achieves the best overall performance, with consistent gains over smaller dimensions, while EM-256 slightly degrades on ACT-4, suggesting that overly large embedding spaces may hurt generalization on smaller datasets.

Table 3. Ablation on embedding size during self-supervised training.

Data	EM-32	EM-64	EM-128	EM-256
ACT-4	53.47	**56.25**	**56.25**	51.39
BIL-6	84.42	85.71	**86.20**	85.71
JML-4	65.49	65.49	67.26	**69.03**

Batch Size. We evaluate batch sizes $\{64, 128, 256\}$ under the same joint training protocol. BS-128 yields the best results across all three datasets, as larger batches provide more within-batch negatives for contrastive learning, while BS-256 shows slight degradation, likely due to reduced gradient diversity at our dataset scale.

Table 4. Ablation on batch size during self-supervised training.

Data	BS-64	BS-128	BS-256
ACT-4	53.47	**56.25**	54.17
BIL-6	83.12	**86.20**	83.12
JML-4	64.60	67.26	**68.33**

Projection Head. We compare an MLP projection head

against a single linear layer. The MLP consistently outperforms the linear layer across all datasets, with the most notable gap on ACT-4 (56.25% vs. 52.08%), confirming that non-linear projections better capture the complex morphological feature interactions needed for effective contrastive representation learning.

Table 5. Ablation on projection head architecture: MLP vs. single linear layer.

Data	MLP	One Layer
ACT-4	**56.25**	52.08
BIL-6	**86.20**	84.42
JML-4	**67.26**	63.72

3. Cross-Domain Generalizability: Neuron-to-Glia Transfer

To assess whether GraPHFormer learns representations that generalize across cell classes, we conducted cross-domain transfer experiments between neuronal and glial morphologies. We obtained glia reconstructions from NeuroMorpho.Org [1] spanning four species: mouse (7,000 samples), rat (2,149), semipalmated sandpiper (1,784), and semipalmated plover (992). Our evaluation protocol consisted of two transfer scenarios: (1) training on neuronal morphologies and testing on glia for species classification, and (2) training on glia and evaluating across six established neuronal benchmarks (BIL-4, JML-4 , ACT-4, N7, M1-Cell, M1-REG). All models were trained for 50 epochs using self-supervised learning and evaluated with frozen k-NN classification following the TreeMoCo protocol.

Table 6. Cross-domain transfer performance between neuronal and glial morphologies. **Left column**: Model trained on neuron data, tested on glia (species classification). **Right columns**: Model trained on glia, tested on neuron datasets (cell-type classification). Results demonstrate transfer capability despite morphological differences between cell classes.

Test Dataset	**Train: Neuron**	**Train: Glia**
Glia (species)	78.87	86.94
BIL-4	-	81.82
JML-4	-	68.14
ACT-4	-	46.52
N7	-	82.78
M1-REG	-	71.08
M1-Cell	-	72.84

Table 6 reveals cross-domain transfer despite morphological differences between neurons and glia. When trained exclusively on neuronal data, GraPHFormer achieves 78.87% accuracy on glia species classification—8 percentage points below the glia-trained model (86.94%). This demonstrates that multimodal topological-structural features transfer across different cell classes.

The reverse transfer—glia-to-neuron—yields competitive performance across several neuronal benchmarks. On N7 (82.78%) and BIL-4 (81.82%), the glia-trained model approaches within-domain performance, suggesting that branching topology and radial geometry encode organizational principles that are shared across cell classes. Performance on JML-4 (68.14%) and cortical datasets (M1-REG: 71.08%, M1-Cell: 72.84%) remains competitive, though the lower ACT-4 result (46.52%) suggests that cortical layer discrimination may benefit more from neuron-specific features.

These findings indicate two properties of GraPHFormer's learned representations: (1) the multimodal fusion of persistence images and graph structure captures morphological signatures that generalize across cell types, and (2) self-supervised contrastive learning discovers features that maintain utility across domain shifts. The ability to transfer between neurons and glia—morphologically distinct yet topologically related—suggests potential for few-shot learning scenarios and cross-species comparative studies where labeled data is limited.

4. Embedding Space Visualization

To qualitatively assess the learned representations, we visualized the GraPHFormer embedding space using t-SNE dimensionality reduction. Figure 3 displays the concatenated multimodal embeddings (tree + image encoders) for four representative datasets: ACT-4 (cortical layers), JML-4 (cortical layers and thalamic neurons), Neuron7 (diverse cell types), and Glia (cross-species). The visualizations reveal clustering patterns that reflect the morphological organization captured by our framework.

For the **ACT-4 dataset** (cortical layer classification), neurons from layers 2/3, 4, 5, and 6 form partially overlapping clusters with some intermixing, particularly between layers 5 and 6. This overlap aligns with the challenging nature of layer-based classification (59.1% accuracy in self-supervised setting) and reflects gradual morphological changes across adjacent cortical layers rather than discrete boundaries.

The **JML-4 dataset** exhibits clearer separation, with VPM (ventral posteromedial thalamic) neurons forming a distinct cluster on the left, while cortical neurons (layers 2/3, 5, 6) show moderate overlap in the center and right regions. This separation pattern is consistent with the higher classification accuracy (72.7%) and demonstrates that GraPHFormer distinguishes between thalamic and cortical projection patterns, which exhibit more pronounced morphological differences than intra-cortical variations.

Neuron7 shows well-organized clustering, with seven cell types forming separated, cohesive groups. Bipolar

and amacrine cells (left), basket cells (center), and pyramidal/spiny neurons (right) occupy distinct regions of the embedding space. This organization (83.8% accuracy, ±0.6% variance) indicates that multimodal fusion captures morphological signatures that distinguish functionally diverse neuronal classes.

The **Glia dataset** (species classification) presents substantial overlap among mouse, rat, and two bird species (semipalmated sandpiper and plover). Despite this overlap—reflecting conserved glial morphologies across species—GraPHFormer achieves 86.94% accuracy, indicating that the learned representations capture species-specific variations in glial process organization that are not immediately apparent in the 2D projection.

These visualizations demonstrate that GraPHFormer learns embedding spaces where morphologically similar cells cluster together, with separation quality correlating with both biological distinctiveness and quantitative classification performance.

Table 7. Ablation study of fusion strategies within the TreeMoCo framework for neuron morphological analysis. All variants employ ResNet-18 as the image encoder and TreeLSTM as the tree-structured encoder, jointly trained on the ACT-4, JML-4, and BIL-6 datasets using contrastive learning. Performance is evaluated on held-out test sets via frozen k-NN classification. We report average classification accuracy (in %) ± standard deviation across three independent random seeds for each fusion method.

Fusion Strategy	ACT-4	BIL-6	JML-4
Bi-Attention	54.39 ± 3.6	87.04 ± 0.7	73.59 ± 4.5
Addition	57.54 ± 4.2	84.5 ± 1.1	72.73 ± 1.3
CAMME [3]	52.28 ± 1.6	84.57 ± 2.1	71.86 ± 1.9
Concatenation	57.89 ± 1.1	83.33 ± 1.8	71.86 ± 1.9
Gated	58.25 ± 4.26	85.19 ± 0.7	72.30 ± 1.9
MHCA	58.90 ± 1.82	84.57 ± 1.07	67.10 ± 0.7

5. Cross-Modal Retrieval Analysis

To evaluate the alignment between tree and image representations, we performed bidirectional cross-modal retrieval experiments. For each query from one modality, we retrieve the top-5 nearest neighbors from the other modality using cosine similarity in the learned embedding space. Figure 4 shows retrieval results for ACT, BIL, and JML-4 datasets in both directions.

Asymmetric similarity scores. Image-to-tree retrieval consistently achieves higher cosine similarities (0.80-0.90 across datasets) compared to tree-to-image retrieval (BIL: 0.12-0.73, JM: 0.00-0.46, ACT: 0.13-0.39). This asymmetry stems from differences in information content: tree graphs preserve detailed geometric information (coordinates, radii, exact connectivity), while persistence images compress morphology into topological summaries through filtration and Gaussian smoothing.

Information bottleneck. When querying with a tree, the detailed geometric specification may not find exact matches in the compressed persistence image space, resulting in lower similarities. Conversely, when querying with a persistence image, multiple geometrically distinct trees sharing similar topological features can match the query pattern, yielding higher similarities.

Semantic alignment preserved. Despite lower absolute scores in tree-to-image retrieval, retrieved neighbors generally belong to the correct morphological class, as evidenced by consistent patterns in persistence images and similar tree structures. This indicates successful alignment through contrastive learning, even with modalities of different information density.

These results demonstrate that both modalities contribute complementary information—graphs provide geometric precision while images offer topological invariance—supporting the effectiveness of our multimodal fusion approach.

6. MoCo-Style Training

We adopted the TreeMoCo framework [2] to evaluate MoCo-style training of our multimodal approach. We integrated the tree encoder and image encoder through various fusion strategies, including multi-headed cross-attention (MHCA), bi-directional cross-attention, addition, concatenation, gated fusion, and CAMME [3]. Our experimental setup and loss function closely follow TreeMoCo, where the objective is to maximize similarity between positive pairs while minimizing similarity between negative pairs.

We employed ResNet-18 as the image encoder and TreeLSTM as the tree encoder. Following TreeMoCo, we jointly trained the models on the BIL-6, ACT-4, and JML-4 datasets in a self-supervised learning scheme. Model performance was evaluated using frozen k-NN classification, with $k = 20$ for BIL-6 and ACT-4, and $k = 5$ for JML-4, consistent with the TreeLSTM evaluation protocol [2]. All models were trained for 50 epochs using SGD with an initial learning rate of 0.06 and batch size 128.

Table 7 presents the fusion strategy ablation results. Performance varies across datasets, with no single strategy dominating universally. On BIL-6, bi-directional attention achieves 87.04% (±0.7), while MHCA performs best on ACT-4 (58.90% ±1.82). Bi-directional attention performs well on JML-4 (73.59% ±4.5), though with notably higher variance. Simple strategies like addition and gated fusion provide competitive results across all datasets, with addi-

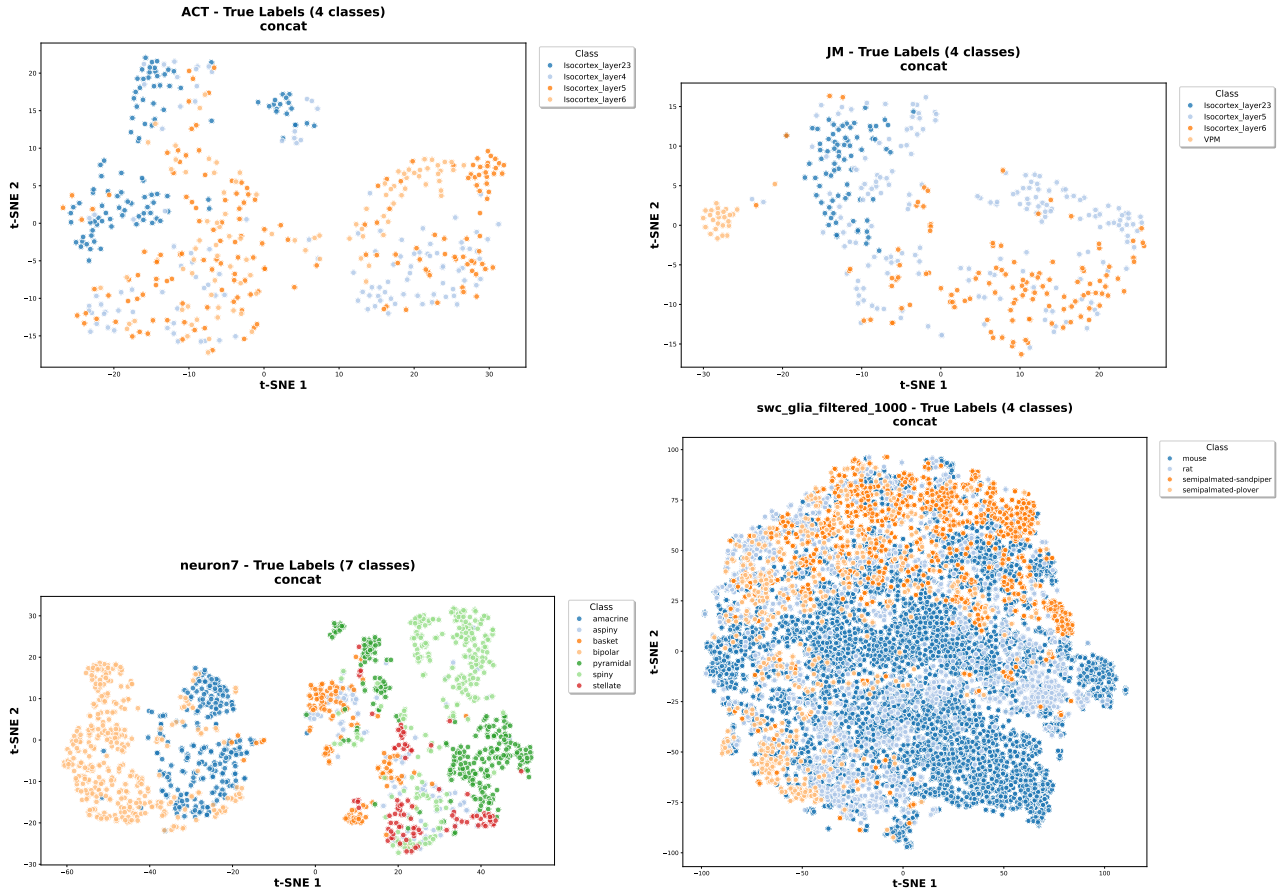

Figure 3. t-SNE visualization of GraPHFormer embedding spaces across four datasets. **Top left**: ACT-4 (cortical layers 2/3, 4, 5, 6) shows partial overlap reflecting morphological gradients between adjacent layers. **Top right**: JML-4 displays separation between thalamic VPM neurons (left cluster) and cortical projection neurons (center-right). **Bottom left**: Neuron7 exhibits distinct clusters for seven morphologically diverse cell types (bipolar, amacrine, basket, pyramidal, spiny, stellate), demonstrating discrimination of fundamental neuronal classes. **Bottom right**: Glia species classification (mouse, rat, semipalmated sandpiper/plover) shows overlapping distributions reflecting conserved morphological features across species. Embeddings are obtained by concatenating tree encoder and image encoder outputs after self-supervised contrastive pretraining. Clustering quality correlates with biological distinctiveness and classification accuracy.

tion achieving 57.54% on ACT-4 and gated fusion reaching 58.25%. These results indicate that fusion strategy selection depends on specific task characteristics, with attention-based mechanisms generally showing competitive performance but with dataset-dependent effectiveness.

7. Visual Correspondence Between Modalities

To illustrate the relationship between tree graphs and their persistence image representations, Figure 5 shows representative examples from ACT, JML, and Glia datasets with tree structures overlaid on their corresponding persistence images. Each row displays samples from a different morphological class, with the tree graph (green) superimposed on the multi-channel persistence image encoding.

The visualizations reveal how morphological features manifest in both representations. For ACT cortical neurons, layer-specific differences in dendritic complexity and branching patterns produce characteristic persistence signatures with varying radial extent (red channel intensity) and branch persistence (green channel). JML-4 samples demonstrate contrast between cortical projection neurons with complex dendritic arbors and thalamic VPM neurons with simpler, more compact morphologies. Glia morphologies demonstrate species-specific variations in process organization, reflected in the spatial distribution and intensity patterns across persistence channels.

The correspondence between tree complexity and persistence image structure confirms that our multi-channel encoding compresses topological information while preserving class-discriminative features. Within-class consistency (similar patterns across rows) and between-class variation (distinct signatures across rows) validate that both modalities capture biologically meaningful morphological differ-

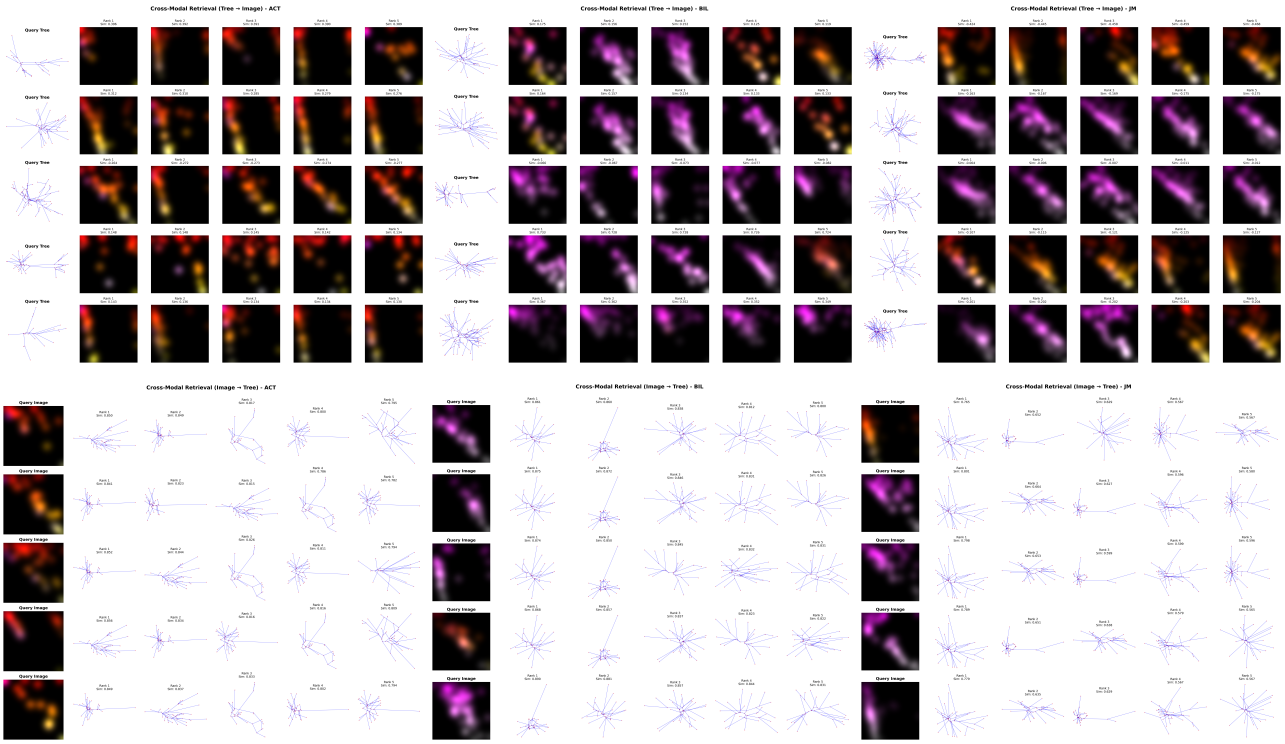

Figure 4. Bidirectional cross-modal retrieval demonstrating alignment between tree graphs and persistence images. **Top row**: Tree-to-image retrieval for ACT, BIL, and JML-4 datasets showing query tree graphs (left) and top-5 retrieved persistence images with cosine similarity scores. **Bottom row**: Image-to-tree retrieval showing query persistence images (left) and top-5 retrieved tree graphs. Note the asymmetry in similarity scores: image-to-tree retrieval achieves higher similarities (0.80-0.87) compared to tree-to-image retrieval (0.10-0.46), reflecting differences in information density between the graph and compressed topological representations.

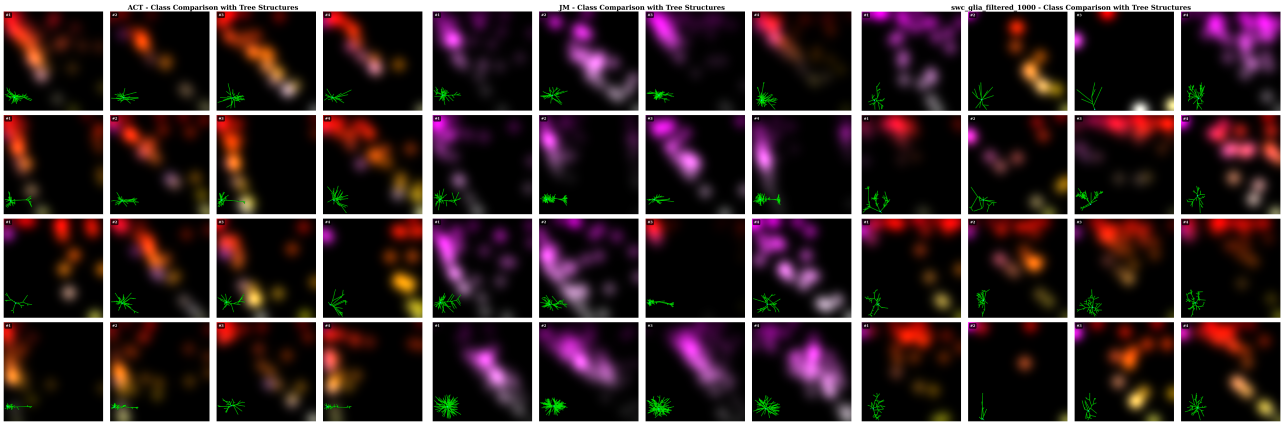

Figure 5. Representative examples showing tree graphs overlaid on their persistence images across different morphological classes. **Left**: ACT dataset (4 cortical layers). **Center**: JML-4 dataset (cortical layers and thalamic VPM neurons). **Right**: Glia dataset (4 species). Each row represents a distinct class, with 4 samples per class. The green tree structures illustrate how branching topology and spatial extent map onto the RGB persistence image encoding, where red captures unweighted density, green encodes persistence-weighted features, and blue represents radius-weighted information.

ences used by GraPHFormer for classification.

References

[1] Giorgio A Ascoli, Duncan E Donohue, and Maryam Halavi. Neuromorpho. org: a central resource for neuronal morpholo-

gies. *Journal of Neuroscience*, 27(35):9247–9251, 2007. 3

[2] Hanbo Chen, Jiawei Yang, Daniel Iascone, Lijuan Liu, Lei He, Hanchuan Peng, and Jianhua Yao. Treemoco: Contrastive neuron morphology representation learning. In *Advances in Neural Information Processing Systems*, pages 25060–25073. Curran Associates, Inc., 2022. 4

[3] Naseem Khan, Tuan Nguyen, Amine Bermak, and Issa Khalil. CAMME: Adaptive deepfake image detection with multi-modal cross-attention, 2025. To appear in Proceedings of the 7th ACM International Conference on Multimedia in Asia (ACM MMAsia 2025). 4